DOCUMENT RESUME

ED 336 967 FL 019 522

AUTHOR Garcia, Peter A.
TITLE The Competency Testing Mine Field: Validation, Legal and Ethical Issues with Implications for Minorities.
PUB DATE 87
NOTE 23p.; In: Theory, Research and Applications: Selected Papers from the Annual Meeting of the National Association for Bilingual Education (16th, Denver, Colorado, March 30-April 3, 1987); see FL 019 511.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Achievement Tests; Admission Criteria; *Competence; Court Litigation; Elementary Secondary Education; English (Second Language); Ethics; Graduation Requirements; Higher Education; Legal Problems; *Minority Groups; *Standardized Tests; *Teacher Certification; *Teacher Qualifications; Test Bias; *Testing Problems; Test Use

ABSTRACT

A genuine concern for the condition of education in the United States has led to misuse of standardized tests to measure teacher quality. The tests have been used for purposes of admission, advancement, merit, graduation, and certification contrary to research findings and test developer guidelines. The current use of tests has resulted in exclusion of ethnic minorities from the mainstream of society. Flawed processes used in test validation have resulted in gross misapplication of standardized tests with serious consequences for American education. Competency testing does not improve the caliber of teachers entering the profession, but simply denies access to some, mostly minorities. In multilingual populations, testing in English discriminates against individuals with other native languages. Remediation for the tests is difficult to accomplish and may not be effective. Court litigation has addressed two well-established legal principles relating to teacher testing: (1) the right of individuals to engage in gainful occupations of their choice, and (2) the state's authority to reasonably regulate. Other measures of teacher success may be more appropriate than competency testing. A 51-item bibliography is included. (Author/MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.

THE COMPETENCY TESTING MINE FIELD: VALIDATION, LEGAL AND ETHICAL ISSUES WITH IMPLICATIONS FOR MINORITIES

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Peter A. Garcia

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY GARCIA, PETER TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED336967

A genuine concern for the condition of education in the United States has lead to the misuse of standardized tests to measure quality. Standardized tests have been used for purposes of admission, advancement, merit, graduation and certification contrary to research findings and test developer guidelines. The current use of tests has resulted in the disastrous exclusion of ethnic minorities into the mainstream of society. Flawed validation processes used in test validation have resulted in gross misapplication of standardized tests with serious consequences for the U.S.

There is a genuine concern on the part of the U.S. people about the preparation and effectiveness of teachers in the Nation's schools. In response to this concern, the current teacher reform movement continues to escalate at a rate that challenges the extent to which a complete monitoring system can be devised. The national concern about the condition of public education has generated steadfast support for accountability in the education establishment.

Innumerable studies and surveys adequately describe the condition of education in America offering a wide range of probable solutions (Peterson et al., 1986). The increased use of standardized tests for assessment purposes and as a means to filter out incompetence from the teacher ranks, has resulted in a bonanza for test developers as the quest for quality education continues. Since 1980, the numbers and varying types of examinations developed, redesigned or modified, or relabled that are in current use, have multiplied in a competitive supply and demand market which is suspect of seeking to exact profits and political gains in the name of educational excellence. The rapid increase in test use has not abated even though researchers and text developers have consistently cited the limited capacity of standardized tests to predict on-the-job performance. This is indicative of the misconception the general public and professionals have about the accuracy of assessment of current tests and their potential to predict how someone will perform in a selected occupation.

FL019522

Legally accepted strategies used by test validators to satisfy validation criteria on standardized examinations have been criticized sharply by

researchers because of how tests are used (Garcia, 1986). Test scores have been used improperly as sole criteria to prevent those who score low on examinations from admission into career professions and occupational preparation pipelines. Findings consistently reveal serious discrepancies between what tests are designed to measure and their ultimate use. The inconsistent relationship between ineffective test validation strategies and the actual use of test scores makes costly validation efforts a questionable practice. For example, misuse of a test can be illustrated when a test would be designed to measure basic skills, then used to evaluate a football coach whose success rests with winning football games. There is no denying the importance of a football coach with an earned teaching certificate having a need for basic skills. The contradiction is that we are in a period when coaches are being dismissed due to low scores on basic skills tests despite their very successful winning careers. The statistical relationship between a coach's success in games won and a designated score on a given basic skills test is not verifiable. Paper and pencil examinations cannot measure the relationships between scores achieved on a test and teaching effectiveness. The literature (Berliner et al., October 1986) on the effective or expert teacher is void of the use of test scores as determinants of competence. Standard test validation criteria used to insure maximum useful interpretation of test results appear in the Standards for Educational and Psychological Tests (1974), and Standards for Educational and Psychological Testing (1985). When the appropriate combination of professionally accepted validation criteria are utilized to develop and validate examinations, inferences made about applicability and appropriate use become more acceptable and accurate. Subsequently, they should be used in combination with other evaluative criteria. The current mandate to develop and use tests, in an effort to improve educational quality, exceeds existing safeguards to protect the integrity of test results and sound test validation practices.

One way of validating examinations for content validity has been through the process of expert judgment. Test questions are reviewed by selected individuals identified as experts in a testing domain. Experts are asked to make judgments about how relevant to the testing domain is a question. The consensus of experts signals that a specific test item is related to the domain being tested and that it can become part of the test. Expert judgment is also used for other purposes, such as setting cut off scores on examinations based on the results of pilot studies (Popham, 1983).

The use of expert judgment for the purposes stated above is not supported by existent research. The writer used the expert judgment method to determine performance levels of beginning first graders in his doctoral

dissertation. Master first grade teachers were unsuccessful in making performance judgments on carefully selected developmental tasks by Gessell. Even at this basic level of performance, true experts in the performance of beginning first graders were unable to make accurate determinations about their performance. It is impossible to consider making accurate judgments about performance on complex examination questions as currently practiced in establishing content validity (Garcia, 1971).

Test developers, psychometric and test experts, are aware of the inferential nature of interpreting test scores with their limited applicability and use. The practice of using unsuspecting experts to make judgments about performance levels, admissions, cut-off scores, hiring practices, merit increases, graduation, certification or tenure is unreasonable, unreliable, unprofessional, unethical and illegal for the following reasons:

A. Making judgments about performance for exacting purposes is not supported by research.

B. Those that are identified as experts are not always selected on their established expertise in the specific domain for which they are making judgments.

C. Questions submitted to the experts are often flawed because they were taken from other examinations and from unrelated sources.

D. Persons identified as experts are often asked to make judgments about cut-off scores without adequate knowledge or awareness of their involvement. These experts often end up designing their own weapon of demise.

E. No evidence exists that supports a demonstrated relationship between test scores and on-the-job performance.

F. The negative impact that these decisions are having on the professional lives of serious students and especially ethnic minorities is unethical.

G. Court decisions made in support of questionable validation processes have resulted in unfair decisions that are contrary to the Constitution and long-established A.P.A.

BEST COPY AVAILABLE

professional standards used to validate tests.

H. The political and economic gains made by test developers and politicians seeking reelection on the back of the educational excellence movement are unprofessional and unethical.

Content Validity

The principle form of test validity used by validators on tests such as the Pre-Professional Skills Test (PPST) is content validity. The limited acceptable use of the content validation process which results in test scores used to make professional decisions about professional careers and people is well recognized by psychometric experts. Inference made from test scores about actual on-the-job performance signals unethical practices which those who develop and sell tests must address. Those who buy standardized examinations are too often unaware of the inability of test scores to measure complex human abilities necessary for exemplary performance in all occupations. The College Board, Education Testing Service and other test developers have ineffectually cautioned test users about the improper interpretation and potential misuse of test scores. The experts agree that content validity is not enough (Fyans, Jr., Garcia 1986) to justify the current use of test scores.

Standardized tests favor native-born white American children (Williams, 1970). The two main problems of standardized testing involve scientific and ethical decisions. Standardized tests are inadequate measures of the capabilities of minorities (Messic & Anderson, 1970). A culture free test of general learning ability is a practical idea, but such test has simply not been developed (Young, 1975). According to Woolever (1985), standardized rating forms used for teacher evaluation are assumed to measure elements of quality teaching and to be reliable. "This is clearly not the case" (p. 23). The accurate measurement of teaching performance is beyond our current ability to measure (Soar, Medley, & Coker, 1983; Sykes, 1983; Zahorik, 1984). Structure (1985) has indicated that:

> While tests of numerical and verbal proficiency can act as a gross tripwire, they do not identify the pragmatic and problem solving skills required by excellent teachers, or merely by competent teachers. Effective teaching is highly contextual and conditional. At its root, teaching is a problem-solving activity. Teachers must make judgments

> and use discretion to develop situationally effective responses to student needs. State requirements that all teachers, despite experience or acknowledged expertise, submit to such blunt and questionable "screening" instruments only can contribute further to an erosion of teachers' sense of status, and promote resentment among those whose cooperation is essential to the prospects of education reform (Structures, 1985).

The validation procedures used on the Arkansas Teacher Test were flawed according to Janet Collins of IOX Assessment Associates, as cited: "For one, field testing of exam items was insufficient because not enough teachers could be coaxed to take practice tests. A copy of the test was circulated among some teachers the night before the exam, and a computer glitch sent score reports to the wrong teachers, said Collins. Further, essay topics were easy to predict, and some teachers wrote "canned" responses. (Education USA, May 25, 1986, p. 280.)"

Yalow and Collins (1985) state, "The processes that can be used to help ensure that these assessment procedures, often paper and pencil tests, possess content validity and, subsequently, to demonstrate this fact, are not yet standardized" (p. 3). Various processes have been used by test developers to successfully stand off legal challenges regarding content validity of tests. Test items are reviewed by a select population of experts who are asked for an opinion on a series of given skills and their applicability to acceptable performance. A belief that teachers should be able to do something is not sufficiently powerful for inclusion as a skill. The skills/knowledge assessed must be necessary for job performance in order to hold up under scrutiny of content validity (Garcia, 1986).

Another method used to demonstrate content validity is the "51 percent rule" used by ETS in South Carolina, North Carolina, and Mississippi. If over 50% of respondents indicate that an item is valid, it is considered to be content valid (Yalow & Collins, 1985). However, since employers cannot justify employment tests on the grounds of content validity if they cannot prove that the content of the examinations includes critical and substantial parts of the job, this procedure is questionable. Any one method used to substantiate content validity of a test is not defensible. There should be an accumulation of sources in order to make a claim of content validity more or less plausible. If a test is given as a screening device, it must reflect knowledge learned, such as that learned in a teacher education program (Yalow & Collins, 1985).

The significant role that cut-off scores play in the testing scenario cannot be underestimated. Cut-off scores are the gatekeepers that select those who may succeed or are deleted from numerous professions. Cut-off scores on identical examinations used in different states may vary considerably from state to state, even though they may be used for identical purposes. Some state users, which are often state education agencies, set their cut-off scores on examinations before they know what their impact might be on the future availability of professionals due to population reductions caused by the tests. States such as Louisiana and New Hampshire have had to lower their cut off scores on the PPST and NTE tests because serious teacher shortages were created by examination cut-off scores. Other states such as Texas and Florida are experiencing serious teacher shortages due to unrealistic testing practices, forcing them to import about 75% of their teachers. The knowledge and skills required to perform successfully in most professions are unmeasurable with pencil-and-paper tests. The practice of using different cut-off scores on identical tests will become a major problem in states practicing reciprocity. The promise of a national examination for teachers by the American Federation of Teachers (AFT), National Education Association (NEA) and other professional organizations bring into sharp focus a permanent problem of national educational conformity. Warner (1986) points out that each individual state controls teacher education and not national accrediting agencies, the federal government, the Carnegie Forum Task Force or the Holmes Group. How will national examination measure state controlled and developed education programs that are different?

The national trend toward increased testing is influenced heavily by test developers, public officials and policymakers. For example, a $20.3 million, five year grant from National Assessment of Educational Progress (NAEP) to Educational Testing Service, a test developer, places federal funds in agencies that make money from selling tests which may be a questionable practice. Unknowingly, the public and policymakers are interested in test scores which they believe are honest indicators of examinee performance. Even educators at the highest levels are unaware or unconcerned about the limited use of test scores. An Education Department representative, Chester Finn, (1986) says that "fairly simple test scores of one kind or another are what will be used, rightly or wrongly, to see if American Education is getting better or worse" (p.14).

The National Picture

The work of J. T. Sandifur (1986) which provides an overview of the types of tests which are reportedly being used in all fifty states stands as a descriptive monument to the testing movement. He has surveyed the nation through state department of education certification officers and directors of testing, and select deans of education. The information gathered is very helpful to those needing comparative information about specific tests being administered and their identified purpose.

There is a natural flaw in interpreting the data reported (by Sandifur) because state departments report the mandated purposes for which tests are being used without demonstrated awareness of the examination's actual testing domain. States that use tests to measure domains for which they were not intended report their mandated purpose and consequently appear in his survey as being used to measure a domain for which the test may not have been designed. The survey by Garcia (1985) showed a clear discrepancy between what some state tests claim to measure and domain actually measured by the reported tests. Some states could not provide adequate validation results or accurate data which would match an appropriate test with a testing domain. Examination use is often inconsistent with test content, test difficulty, and test use.

The discontentment of public educators whose jobs are threatened with the testing mandate has caused politicians who require examinations to retreat from their mandates. The broad discontentment of teachers which has become particularly evident in states such as Texas has resulted in the reversal of legislated testing of teachers in the content areas. Former promises by the governor which would require urine tests for teachers, quickly became sympathetic pleas to eliminate tests for teachers as the November elections come into view. The dreaded teachers' test in Texas, known as the Texas Examination of Current Administrators and Teachers (TECAT), caused months of worry to discover later that over 99.9% of white test takers passed the test. The millions spent on test development only created a morale problem that has served notice on politicians that teachers would exercise their voting power.

The inequities that have resulted from the testing movement are continually cited by researchers and proponents of equity. The women's movement has propelled this group into job market opportunities by well organized women's political organizations during the past 10 years (USA

TODAY, 1986). Conversely, the numbers of ethnic minorities which have always been seriously underrepresented, are experiencing devastating setbacks with the testing movement. Blacks have been represented at higher rates than Hispanics in public education. The writer's projections show that the testing movement, if continued unabated, is reducing the numbers of ethnic minorities (Blacks and Hispanics) from serious underrepresentation to dismal exclusion (Garcia, 1986). Pressman and Gartner (1986, p. 13) claim that tests that are used in the name of "educational excellence" are simply a code word for removal of Blacks and Hispanics from the teaching profession

Teacher education testing is having a devastating effect on ethnic minorities (Anrig, 1985; Brown, 1985; Ekstrom & Goertz, 1985; Goertz & Pitcher, 1985; Popham, 1984; "Teacher exams," 1985). The pass rates of ethnic minority groups on tests for entry into teacher education programs, credentialing, and certification continue to restrict the numbers of ethnic minorities entering and remaining in teaching (Goertz & Pitcher, 1985). Teacher shortages in certain types of inner city school districts have already been noted. Teacher shortages will continue to be a problem for ethnic minority inclusion in the nation's teaching corps (Ekstrom & Goertz, 1985; Smith, in press).

A significant step was taken in California to insure greater access to Hispanics into higher educational system. Chancellor W. Ann Reynolds of the California State University reports that although 25% of the state's school enrollment is Hispanic, only 10% of the university system's students are Hispanic ("Ca. Panel Urges", 1985). She indicates that, "California cannot afford this massive and growing loss of talent" (p. 2). Chancellor Reynolds appointed a Commission on Hispanic Underrepresentation, a panel of the 19 campus California State University system, in May 1984. The commission has recommended 46 ways to increase the enrollment of Hispanic students. "Some of the recommendations involve major programs with substantial costs; others are low in cost and/or more narrow in scope" (Arciniega & Morey, 1985, p. 37). Three of the recommendations are as follows:

1. Programs to help high schools with large Hispanic enrollment to strengthen their college-preparatory courses and improve counseling.

2. The assignment of state university faculty members as "mentors" for Hispanic high school students.

3. Programs to help Hispanics succeed on college entrance examinations. (Cal. Panel Urges, 1985, p.2.)

According to Arciniega and Morey (1985), "The Commission has identified the following areas in which immediate action is necessary: 1) improved access to college, 2) direct services to students to assist them to progress educationally, and 3) the enhancement of institutional capabilities" (p. 37).

The poor performance of minorities on competency tests, for certification, have been well documented (Lindahl & Wholeben, 1985). The mean scores of whites is 18 to 20 points higher than for Blacks on all three parts of the NTE. Hispanics scored 0.6 to 0.8 standard deviations below the mean for whites. The current trend will reduce the heterogeneity of the teachers in America (Anrig, 1985; Ekstrom & Goertz, 1985; Goertz & Pitcher, 1985). Passing rates of Anglos are consistently higher across the areas of math, reading and writing.

The Washington Post (Richburg, June 28, 1985) reported that 10% of the teachers failed the statewide competency test given to teachers in Arkansas. Out of 28,000 public school teachers, 2,803 failed the test. About seven percent failed the writing portion, five percent failed the reading, and three percent failed the math. In the predominantly black Lee County, 34.5% of the teachers failed the test, while in Carroll County, with a mostly white population, 2.6% of the teachers failed. "When Louisiana decided to require teacher education candidates to pass the National Teacher Examinations, Grambling State University found itself where many black colleges do - at the bottom. Fewer than 10% of its students qualified" (Up From the Bottom, 1985, p. 238).

Impact on Teacher Education Programs

Closing teacher education programs in institutions with open admission criteria which serve primarily minority populations will worsen the socioeconomic and racial/ethnic mismatch between students and teachers. The current teaching profession is 87% white, 10% black and 2% Hispanic. States having large ethnic minority populations, such as California, New York, Texas and Arizona, have current testing programs which assure a decline of those minorities entering colleges with aspirations of becoming teachers (Goertz & Pitcher, 1985).

Ekstrom & Goertz (1985), Goertz & Pitcher (1985), Goertz, Edstrom and Coley (1984), Smith (in press), and Kidd (1984) indicate that testing policies in most states have been mandated by legislators responding to the "excellence in education" movement. A review of actions taken by legislative bodies suggests that state policies are shortsighted and unaware of the impact they will have on the teacher supply and on ethnic minority representation in American education (Ekstrom & Goertz, 1985). Testing continues to be the cheapest way to evaluate teachers, but is not likely to improve the caliber of those wishing to enter the profession. Instead, it simply denies access to those who score low on competency examinations (Teacher Exams, 1985). Peterson (1984) from ETS makes an important point which concerns many educators across the country who are dedicated to honesty in test application: "I consider much teacher testing policy to be badly conceived, hastily and/or mindlessly conceived, put forward for the wrong reasons, and then often finally shaped to the interests of the strongest power blocks. All of this, needless to say, is not unique to education policy-making. Someone once said there are two things you don't ever want to see being made: One is sausage, the other is law. (p. 15)"

Impact On Multilingual Populations

Hispanic-Americans make up the largest ethnolinguistic group in America. The disparity in the performance of Hispanics and Whites on standardized achievement tests has long been known. The findings of Friedman (1985) confirm the fact that students tested in their dominant language score higher than when tested in their non-dominant language. The findings of other researchers verify language as a factor in statewide testing of ethnic populations.

Low social class is often associated with ethnic minorities. According to Campbell, Cunningham, Usdan, & Nystrand (1980): "The correlation between social class and academic success is well known. Students from middle and upper class backgrounds are more likely than lower class students to do well on achievement tests, graduate from high school, and go on to college. On the other hand, lower class students are more likely to enter school with severe educational deficiencies, behave in ways which middle class teachers consider inappropriate and leave school at an early age (p. 308)."

Test scores for the CBEST (The original tests were clones of the

PPST.) in California revealed the same problem of higher failure rates for ethnic minorities. The pass/fail rates were similar in California on the CBEST as in Texas on the PPST, with the exception that the pass rates were even lower in Texas for minorities. This is explained by the different populations who took the test. In California, the early examinees included large numbers of teachers seeking certification, as compared with students seeking admission into teacher education programs in Texas. Students coming from educationally impoverished backgrounds, where English was not spoken, scored lower on the CBEST. In Georgia, students from historically black colleges scored sufficiently low to place teacher training programs at these institutions on probation. In California, teacher candidates scored higher than employed teachers which, perhaps, results from lack of familiarity with the material tested (Peterson, 1984).

Remediation for these tests becomes nearly impossible. The PPST, CBEST, and CAT are difficult to teach because institutions cannot specify deficiencies within each basic skills subtest. There is no evidence to indicate that institutions are coordinating their efforts to provide needed remediation (Ekstrom & Goertz, 1985). ETS also predicts only minor remediation possible based on its data base of PPST test results. Forty percent of the retakes on the PPST fail to change their scores (Goodison, 1985).

Court decisions have systematically been made in support of continued test use without regard for their impact on ethnic minorities. The recent decision in the case of U.S. vs Texas addressed the use of the PreProfessional Skills Test (PPST) and its devastating effect on Blacks and Hispanics. The court ruled in favor of continued use of the PPST based on previous court cases and faulty earlier court decisions.

The court's role in affirming test use and particularly the content validity of the National Teacher Examinations (NTE) has been pointed out by A. J. Wilson (1984b) who states:

> By the 1970's, renewed attention was paid to questions of test content and test validity with the emergence of a number of law suits charging that the tests were being used in some states and communities to discriminate against minority teachers and teacher candidates. In 1978, U.S. Supreme Court, relying heavily on a study conducted for the state of South Carolina by the Educational Testing Service, ruled in favor of the exam's use and thus indirectly in favor of their content validity. The court affirmed a lower court's

> decision which stated that 'the NTE have demonstrated to provide a useful measure of the extent to which prospective teachers have mastered the content of their teacher training program'(p. 2).

The NTE is a standardized examination in which validity relates most directly to the content of teacher education programs. The broad scope of the knowledge tested indicates that the best preparation for the test is the knowledge and experience one gains from a teacher preparation program. The scores are not exact measures, and each score has associated with it a degree of error. NTE policy does not set passing scores for the examinations. ETS offers guidelines for setting minimum standards and test score use. It encourages recipients to use test scores as one of several bases for making decisions about certification and selection of teachers (ETS, 1985). However, Toch (1984) reports an increase in the use of the NTE as an admissions requirement. This has reduced the number Blacks entering the teaching profession.

Olstad, Bean, Foster, & Marrett (1985) did a validity study on the general knowledge component of the NTE Core. They found a rather consistent correlation between the NTE Test of General Knowledge and the California Achievement Test (CAT). They also noted a lack of correlation between the GPA and either the NTE or the CAT which suggests that they measure different factors. A 1983 study by Olstad, Beal, Schlick-Noe, & Schaefer at the University of Washington suggested a predictive value of GPA for teaching performance (cited in Olstad et al., 1985), whereas, the CAT did not show a predictive value. According to Olstad et al., the correlation shown between the NTE and CAT would indicate that the NTE has no predictive value in assessing success in a student teaching program. The NTE has not been empirically correlated with student teaching outcomes. Peterson (1984) cites a correlation reported in Georgia and Oklahoma of .30 and .40 respectively between student teaching and on the job performance.

Legal Aspects

Freeman, Hess, & Kasik (1985) have discussed extensively the legal aspects of teacher testing. The courts have examined two well established legal principles which have impacted on testing across the nation. The first is the right of individuals to engage in the gainful occupations of their choice as protected by due process. The second is the state's authority to reasonably regulate in order to protect the public interest. Griggs v. Duke Power Company (1971) reversed the "common sense standard" used up to that time which relied upon the credibility of testimony or on minimum attempts at validation (Freeman et al., 1985). In Griggs, the "court eventually decided that any test is illegal unless it clearly measures the skills needed for the job" (p. 3). Hunnicut v. Burge (1973) concluded that qualified teachers should possess "a good education."

The decision in Washington v. Davis (1976) modified interpretations given by the courts earlier because intent rather than impact was stressed (Freeman et al., 1985). The court ruled that tests' job relatedness need not be considered unless actual intent to discriminate was shown. This has provided much greater latitude in test interpretation. When a test seems "obviously" relevant and was developed "in good faith," it is accepted and the Equal Employment Opportunity Commission (EEOC) guidelines for job relatedness are not considered (Freeman et al., 1985).

The U.S. v South Carolina (1977, 1978) decision permitted the use of tests to certify minimally qualified persons through the use of a content validated standardized test (Freeman et al., 1985). The court ruled that "standardized tests do reflect individual achievement with respect to specific subject matter content which is directly relevant to (although not sufficient in itself to assure) competence to teach" (cited in Freeman et al., 1985, pp. 25-26). Hazard, Freeman, Eisendorfer, & Tractenberg (1977) stated that a general education is what each teacher should have, giving greater generalizability to test developers (cited in Freeman et al., 1985). Steward v. Hannon (1979) dealt with licensing exams for principals and upheld the use of examinations for this purpose (Freeman et al., 1985).

Mohammed (1984) of Golden Rule Insurance Company raises questions about the appropriateness of the test used in Golden Rule which are similar to those raised with other examinations like the PPST. The examination covers material viewed as inappropriate for an entry-level examination. The test requires high level test taking ability with linguistic

and vocabulary skills unrelated to an applicant's competency and trustworthiness. Cut-off scores were established without sufficient regard for application. The plaintiffs claimed racial discrimination in violation of the Equal Protection Clauses of the U.S. and Illinois Constitution, and a deprivation of due process resulting from the arbitrariness of many aspects of the examination.

The out-of-court settlement of Golden Rule Insurance v. Mathis (1980) in November, 1984, continues to have impact on future testing because the settlement, in effect, requires greater supervision of tests and their uses (Freeman et al., 1985). The settlement which evolved subsequent to the Illinois Supreme Court's denial of defendants' petition to file appeals contains these fundamental points (pp. 11-12).

> (1) Applicants taking subsequent examinations will be asked to furnish voluntarily their race or ethnicity and the level of education they have attained.
>
> (2) The Director of the Department of Insurance will publish annually a report containing statistical information, including the results of the testing by race for each part and the entire test, the mean scaled scores on each part, and the standard deviation of scaled scores on each part, and the results by race for those having a high school diploma or a G.E.D. In addition, the Department is required to prepare an "item report" which provides the correct answer rates for each item by race as well as r-biserial correlations.
>
> (3) In constructing future tests, ETS is obliged to adhere to both the APA's Standards for Educational and Psychological Tests and ETS' Standards for Quality and Fairness and Test Sensitivity Guidelines.
>
> (4) The agreement stipulates that no more than a twelfth grade reading level as determined by generally accepted reading-factor indexes, e.g., FOG or SMOG, shall be required (excluding terms specifically related to the insurance industry, e.g., the term beneficiary).
>
> (5) The construction of future tests, according to the agreement, is to proceed in the following manner:

All test items are to be divided into two types. Type I items are those for which a) the correct answer rates of black examinees, white examinees, and all examinees are not lower than forty percent (40%) at the .05 level of statistical significance and b) the correct answer rate of black and white examinees differs by no more than fifteen percentage points at the .05 level of statistical significance. Type II items are all other items.

The agreement then stipulates that Type I items are to be used exclusively as long as there is a sufficient number. When used, Type II items are to be selected in descending order of the least difference between white and black examinees. New items may be generated, but must be pretested on three occasions; these items, known as Type III, cannot be used until after being pretested and then classified as Type I or Type II items.

(6) Finally, the agreement stipulates the creation of an advisory committee comprised of representatives of the insurance industry in Illinois and two persons knowledgeable in the area of psychometrics. The committee will review test results and make recommendations concerning the test.

Some independent school districts have developed their own tests and validation studies. The Houston Public Academic Skills Test (FAST) assesses basic reading, writing, and mathematics skills. According to W. N. Kirby, Commissioner of Education in Texas, of the 12,000 to 13,000 professional staff from Houston Independent School District, 7,000 have passed the reading and writing parts of the FAST examination (Kirby, 1985). The "face validity" approval given to the FAST by the Texas Education Agency staff appears to be inconsistent with the guidelines of the Standards for Educational and Psychological Testing (1985).

The use of any test for employment selection, promotion, or retention requires that a test be appropriately validated by the agency administering the examination. Beyond this, there are professional, ethical, and legal responsibilities that the examiners must safeguard against should the examination have an adverse impact on hiring, promotion, or membership opportunities of members of any race, ethnic, or gender group. Employees may file suit against an employer, claiming discrimination in employment practice, if criteria and procedures for evaluation are not consistent with the Uniform Guidelines in Employment Selection Procedures, published in 1978 (Stein & Frankiewicz, 1985).

To be legal, a testing program must be equitable and fair. Examinations must measure what teachers should know to be effective teachers. Tests cannot be used to discriminate by race or gender. The safest possible legal ground for public administrators are indicated by Lines (1985) as follows:

1. Appropriate validation of tests.
2. Insure that tests measure what they are intended to measure and that this reflects skills needed for on-the-job performance.
3. Use tests only for purposes and applications recommended by the developer.
4. Avoid tests which disproportionately exclude a racial or gender group.
5. Use additional non-test appropriate criteria.
6. Insure adequate procedures for challenging errors or abuses.

It is important to recognize the differences between present testing criteria and legal criteria used by the courts for dismissal purposes over the past 40 years. The courts have accepted pre-employment tests with some reservations, and require that they be equitable and fair. Seventeen states have adopted a testing requirement for new teachers. Fourteen are considering general competency tests. The courts reportedly are willing to question efforts which are not a reasonable measure of performance (Lines, 1985). Texas approved a state wide test of basic skills to be taken by all teachers in the area of reading and writing (Langford, 1985). The test was first administered in March of 1986 with confusion and organized revolt resulting from a judge's ruling that one teacher was not required to take it (Stolarek, 1986b).

The development and applicability of tests in South Carolina are technically (and thus probably also legally) defensible based on the Angoff, Nedelsky, and Ebel approaches (Hamm & Winer, 1985). These approaches require groups of experts to make item- by-item judgments of each test item difficulty. However, the process of asking experts to make judgments about

actual performance as a method of measuring actual performance is not sufficiently accurate (Garcia, 1971).

The responses to and the reasons for mandating change in the testing movement in this country are developing:

1. In Florida, both teacher unions have filed litigation against the Master Teacher Plan, causing it to be terminated (1986). At least 10 other states have legislation pending or have passed similar legislation (G.W. Wilson, 1985).

2. There is insufficient research evidence favoring one method of job analysis for construction and validation of personnel tests indicating that there should be multiple methods used.

3. Tests developed on the basis of systematic job analysis are more fair than commercially developed tests.

4. Specific written knowledge tests developed through job analysis are more fair to minorities (Stein & Frankiewicz, 1985).

5. Judge Myron H. Thompson ordered major changes in Alabama's Initial Teacher Certification Test in early 1986 and will likely be appealed in the 11th Circuit Court (Judge, 1986).

Summation

It is becoming apparent that the testing movement can no longer be the tail that wags the dog. The joint meeting of NEA, ETS, AACTE, and Fairtest on December 11th and 12th, 1986 in Washington, D.C., is evidence that divergent interest groups must address the question of integrity in testing. Monetary and political interests must lean with the pressures for equity in education with full assurance that quality will not be sacrificed in America. The leadership of these groups agreed that criteria other than standardized tests must be used to measure teaching competence. Dr. Greg Anrig, President of Educational Testing Service reported that the student teaching experience is the best measure of the success of a teacher. That raises a serious question about why standardized tests are used for this purpose.

REFERENCES

Anrig, G.R. (1958b, March). Teacher education and teacher testing: The rush to mandate. Speech presented at the annual convention of the American Association of Colleges for Teacher Education, Denver, CO.

Arciniega, T. A., & Morey, A. I. (1985, June). Hispanics and higher education: A CSU imperative (Final Report of the Commission on Hispanic Underrepresentation). Long Beach, CA: The California State University, Office of the Chancellor.

Berliner, David, & Brandt, Ron et al. (1986). The expert teacher. Educational Leadership, 10, 44, No. 2, pp. 3- 60.

Brown, J. F. (1985, February 6). Implications of basic skills testing for the ethnic group composition of professional school staffs. Memo to members of the Executive Committee of the Commission on Teacher Credentialing, Sacramento, CA.

California panel urges steps to help Hispanics. (1985,July). The Chronicle of Higher Education, p. 2.

Campbell, R. F., Cunningham, L. L., Nystrand, R. O., & Usand, M. D. (1980). The organization and control of American schools (4th ed.). Columbus, OH: Merrill.

Collins, Janet. (1986). Education USA,p. 280.

Educational Testing Service. (1985b). Pre-Professional Skills Tests Score Interpretation Guide. Princeton: NJ.

Ekstrom, R. B., & Goertz, M. E. (1985, March). The teacher supply pipeline: The view from four states. Paper presented at the annual meeting of the American Educational Research Association. Chicago:IL.

Freeman, L. D., Hess, R., III, & Kasik, M. M. (1985, March). Testing teachers and the law. Paper presented at the annual meeting of the American Educational Research Association. Chicago: IL.

Friedman, S. G. (1985, March). Testing language, language dominance, and

math scores of bilingual Hispanic students. Paper presented at the annual meeting of the American Educational Research Association. Chicago: IL.

Fyans, Jr. Leslie J., Chair. Symposium: (1986, August) Criterion Validity in Certification and Licensure Circumstances. American Psychological Association 94th Annual Convention, Washington, D. C.

Garcia, P. A. (1971). Differences between expectations of performance in beginning first graders and actual performance. Unpublished doctoral thesis, Oregon State University, Corvallis, OR.

Garcia, P. A. (1986, Fall). The Promise of a National Test for Teachers. Teacher Education & Practice. 2 (2), pp. 45-46.

Goertz, M. E., Ekstrom, R. B., & Coley, R. J. (1984). The impact of NTE use by states on teacher selection (Research Report No. RR-85-1). Princeon, NJ: Educational Testing Service.

Goodison, M. (1985, March). Testing the basic competencies of teacher education candidates with the Pre-Professional Skills Tests (PPST). Paper presented at the annual meeting of the American Educational Research Association. Chicago: IL.

Hamm, D. W., & Winter, P. C. (1985, March). The education entrance examination: Testing the basic skills of prospective teachers in South Carolina. In A. M. Gallegos (Chair), Basic Skills testing for prospective teachers: Promises, pitfalls, and pratfalls. Symposium conducted at the annual meeting of the American Educational Research Association. Chicago: IL.

Judge in Ala. Reverses order on teacher certification test. (1986, February). The Chronicale of Higher Education.

Kidd, J. L. (1984). Crossing the rubicon: The impact of teacher testing. Teacher Education & Practice, 1 (1), 27-33.

Kirby, W. N. (1985, June). Request for conditional approval of an alternative examination to the Texas examination of current administrators and teachers TECAT. Letter to the Texas State

Board of Education. Austin: TX.

Langford, M. (1985, June 9). Board of education approves teacher testing. The McAllen Monitor(pp. 1A, 9A).

Lindahl, N., & Wholeben, B. (1985). Teacher education candidates and the performance of basic skills. Teacher Education & Practice, 2 (1), 35-42.

Lines, P. M. (1985). Testing the teacher: Are there legal pitfalls? Phi Delta Kappan, 66, 618-622.

Messick, S., & Anderson, S. (1970). Educational testing, individual development, and social responsibility. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 047 003).

Mohammed, K. (1984, November). Golden Rule Insurance Company et al. v. Educational Testing Service and the Director of the Illinois Department of Insurance, Background of Lawsuit. (Available from Kimathi Mohammed, Golden Rule Insurance Company, Golden Rule Building, The Water Front, Indianapolis, IN 46224-4199.)

Olstad, R. G., Bean, J. L., Foster, C. D. & Marrett, A. V. (1985). A validity study of NTE general knowledge component as a predictive instrument for successful student teaching. (Teacher Education Research Center Report, Report No. 85-1.) Seattle: University of Washington, College of Education.

Peterson, Jerry, et al., (1986). Productive School Systems for a nontraditional world. Association for Supervision and Curriculum Development. Alexandria: VA.

Peterson, R. E. (1984, April). CBEST, NTE, and other mensurations: Notes on testing would-be teachers in California and elsewhere. Talk prepared for the Spring conference of the California Council on the Education of Teachers. San Diego: CA.

Popham, W. J., Yalow, E. S., & Appel, E. M. (1983). Appraising the Pre-Professional Skills Test for the state of Texas : Report Number three: Textbook analysis. Prepared for the Texas Education

Agency by IOX Assessment Associates, Culver City, CA.

Popham, W. J. (1984). Teacher competency testing: The devil's dilemna. Teacher Education & Practice, 1 (1), 5-9.

Pressman, H., and Gartner, A. (1986). The New Racism in Education. Social Policy, 33 West 42nd St., New York, New York 10036. 17, No. 1.

Richburg, K.B. (1985, February). Federal exams still unavailable: Affirmativeaction effort is stalled. The Washington Post,p.A9.

Sandifur, J. T. (1985, March). State assessment trends. AACTE Briefs, 6 (2), pp. 21-23.

Smith, G. P. (in press). The impact of competency tests on teacher education: Ethical and legal issues in selecting and certifying teachers. In M. Haberman (Ed.). Research in Teacher Education. Norwood, NJ: ABLEX Publishing Company.

Soar, R. S., Medley, D. M., & Coker, H. (1983). Teacher evaluation: A critique of currently used methods. Phi Delta Kappan, 65, 239-246.

Stein, T., & Frankiewicz, R. G. (1985, March). Functional academic skills tests validation study for inservice teachers in Houston I.S.D. In M. F. Brightman (Chair), Local school district development and implementation of teacher basic skills competency testing. Symposium conducted at the annual meeting of the American Educational Research Association. Chicago: IL.

Stolarek, J. (1986b, March). Teachers: TECAT fight to continue. The McAllen Monitor(pp. 1A, 10A).

Structures to Attract & Retain Teachers. (1985, Fall). IEG Policy Notes, 6 (3),p. 1.

Sykes, G. (1983). Contradictions, ironies, and promises unfulfilled: A comtemporary account of the status of teaching. Phi Delta Kappan, 65, 87-93.

Teacher exams: The new rage... (1985, January). Education USA, 27, pp. 144-145.

Toch, T. (1984, December). Test organization, insurance firm settle bias suit. Education Week, pp. 1, 14.

Up from the bottom. (1985, March). Education USA, 27, p. 238.

Warner, Al. Action in teacher education, The Journal of the Association of Teacher Educators. A professional house divided, Fall, 1986, Vol. VIII, No. 3, p. 21.

Williams, R. L. (1970). Effective teaching: Gauging learning while teaching. The Journal of Higher Education, 56, 320-337.

Wilson, A. J. (1984b). Questions of content validity: The evolving relationship of the National Teacher Examinations to teacher education curricula. Proposal submitted to the American Educational Research Association for the 1985 annual meeting. Chicago: IL.

Wilson, G. W. (1985, April). The Florida Master Teacher program. In W. Loadman (Chair), A critical examination of evaluation procedures in four merit pay/career ladder plans. Symposium conducted at the annual meeting of the American Educational Research Association. Chicago: IL.

Woolever, R. M. (1985). State-mandated performance evaluation of beginning teachers: Implications for teacher educators. Journal of Teacher Education, 36, (2), 22-25.

Yalow, E. S., & Collins, J. L. (1985, April). Meeting the challenge of content validity. In J. Rose (Chair), The assessment boomerang returns: Competency tests for education. Symposium conducted at the annual meeting of the American Educational Research Association. Chicago: IL.

Young, P. A. (1975). A culture free performance test of general learning ability (Master's thesis, Naval Postgraduate School). Springfield, VA: National Technical Information Service.

Zahorik, J. A. (1984). Can teachers adopt research findings? Journal of Teacher Education, 35 (1), 34-36.